Impact of Alphabet Knowledge on Online Writer Identification

Guoxian Tan, Christian Viard-Gaudin, Alex Kot

2009 10th International Conference on Document Analysis and Recognition

Impact of Alphabet Knowledge on Online Writer Identification

Guo Xian Tan[1, 2], Christian Viard-Gaudin[2], Alex C. Kot[1]
[1]*Nanyang Technological University, Singapore*
[2]*IRCCyN, UMR CNRS 6597, Ecole Polytechnique de l'Université de Nantes, France*
tanguoxian@pmail.ntu.edu.sg, christian.viard-gaudin@univ-nantes.fr, eackot@ntu.edu.sg

Abstract

Character prototype approaches for writer identification produces a consistent set of templates that are used to model the handwriting styles of writers, thereby allowing high accuracies to be attained. This paper extends such work on writer identification by investigating the usage of alphabet knowledge derived from the character prototypes. In addition, we demonstrate the concept of discriminative power of alphabets. It is not unconceivable that certain alphabets allow writers to express their individuality of handwriting with a more distinct and unique style compared with other alphabets. This paper establishes that such alphabets have higher discriminative powers in identifying writers. Experiments related to the reduction in dimensionality of the writer identification system are also reported. Our results show that the discriminative power of alphabet can be used to reduce the complexity while maintaining the same level of performance for the writer identification system.

1. Introduction

Writer identification from handwritten documents garners much attention in the research community as this is a field that demands both multi-disciplinary and inter-disciplinary knowledge extending from pattern recognition to behavioral sciences. Handwritings of individuals vary distinctly across different cultures, occupations, physical attributes and even physiological factors as discussed by Huber, Headrick and Morris [1, 2]. Writer identification exploits this individuality of handwriting styles in the analysis of the handwriting to differentiate different writers apart. Knowledge of the identity of the writers is important in applications such as forensic document analysis, biometrics-based authentication systems and even writer-adaptation.

Currently, most of the state-of-the-art in writer identification utilizes a template-matching approach to model the handwriting styles with prototype templates [3-5]. These prototypes are based on some global criteria such as texture, slant and curvature features [3] as well as local criteria such as allographs, graphemes, and connected components [6]. Recent advances in writer identification depict an increasing trend of more Information Retrieval (IR) based prototype matching approaches due to its simplicity in design and encouraging results [3-5, 7, 8]. Bensefia et al. can be credited with introducing the concept of IR to the context of writer identification. They utilized grapheme prototypes to attain an accuracy of 95% on 88 writers who wrote French text and 86% on 150 writers who wrote English text. Bulacu et al. used graphemes as well as textural prototypes to attain an accuracy of 92% on 650 writers who wrote English text. Tan et al. and Niels et al. worked at the character prototype level to attain an accuracy of 99.2% on 120 writers who wrote French text and 100% on 43 writers who wrote English text respectively. The advantage of working at the character level is that more consistent prototypes can be built, thus explaining the improvement in results.

Our work is an extension to the proposed method by Tan et al. in [5]. The originality of this paper lies in our usage of alphabet knowledge as additional clues to assist in the writer identification process. This paper presents evidence to show the impact of alphabet knowledge on writer identification. In addition, results on the discriminative power of different alphabets in identifying writers are discussed in this paper. Finally, we show that the overall complexity can be reduced using the discriminative power of alphabets, without much adverse effect to the performance of the writer identification system.

The remainder of this paper is organized as follows: Section 2 describes the proposed methodology and experimental setup. Following that, section 3 then presents the experimental results. Finally, discussions and future areas to explore are described in section 4.

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.165

2. System Architecture

The writer identification system is based on our previous works [5]. The overall system architecture can be summarized by partitioning it into three steps consisting of the prototype training stage, the document indexing stage and the retrieval stage. The purpose of the prototype training stage is to build a set of character prototypes using the IRONOFF database [9] to model the different handwriting styles. The segmentation of the characters to build the set of character prototypes is done by an industrial text recognition engine, MyScript [10]. The second stage then utilizes these prototypes to transform characters segmented from the reference and test documents, using the same industrial engine, into a distribution of frequency vectors: the term frequency, *tf* and the inverse document frequency, *idf*. We have adapted the standard *tf* and *idf* measures from traditional document analysis literature [11] to address our writer identification problem. In our framework, *tf* and *idf* are used to create a statistical distribution that models the handwriting styles of different writers. This distribution then enables us to classify and perform an identification of the writer in question. A detailed account of the inverse document frequency and term frequency is given in our previous work [5]. Finally, the *tf* and *idf* are then classified to identify the writers in the last retrieval stage. In this paper, we propose to use alphabet knowledge derived from the character prototypes to investigate the impact on the identification rate. Two writer identification systems are built; one that uses alphabet knowledge and one without alphabet knowledge. The experimental setups for both systems are further described in the following sections.

2.1. Methodology without alphabet knowledge

In the prototype building stage, the characters segmented by the industrial recognition engine are clustered into 260 prototypes all at the same time in the feature space. This is accomplished without any alphabet knowledge using the well-established k-means clustering algorithm [12]. Thereafter, in the document indexing stage, a fuzzy c-means approach described by Tan et al. [5] is used to map the characters segmented from the reference and test documents to the 260 prototypes. Hence, the *tf* and *idf* are calculated in this setup without any alphabet knowledge. Classification of the test documents to the reference documents is achieved using equation 1, where N=260 stands for the number of prototypes that are being clustered in this setup.

$$\text{Distance}\left(\text{writer}_i, \text{writer}_j\right) = \sum_{k=1}^{N} \frac{idf_k \left(tf_{k,i} - tf_{k,j}\right)^2}{tf_{k,i} + tf_{k,j}} \quad (1)$$

2.2. Methodology using alphabet knowledge

The experimental setup using alphabet knowledge follows the three main stages described previously, except for one notable exception; the character prototypes are built on an alphabet basis, using alphabet knowledge of the characters during clustering. A choice of a set of 10 clusters for each alphabet is used so that we can make a comparison with the approach described previously. Hence, there is a common set of 26x10=260 prototypes to model the handwriting style of each writer. In addition, we introduced an *Alphabet Information Coefficient* (*AIC*) during classification, as shown in equations 2 and 3.

$$AIC_\alpha = \frac{1 - \exp(-\lambda \times C_{\alpha_i} \times C_{\alpha_j})}{1 + \exp(-\lambda \times C_{\alpha_i} \times C_{\alpha_j})} \quad (2)$$

$$\text{Distance}\left(\text{writer}_i, \text{writer}_j\right) = \frac{\sum_{\alpha=a}^{'z'} \left(AIC_\alpha \times \sum_{k=1}^{N} \frac{idf_k \left(tf_{k,i} - tf_{k,j}\right)^2}{tf_{k,i} + tf_{k,j}} \right)}{\sum_{\alpha=a}^{'z'} AIC_\alpha} \quad (3)$$

In equation 2, C_{α_i} and C_{α_j} represent the number of characters or instances of alphabet α, that appear in the documents of reference writer i and test writer j respectively. Following this, the *AIC* is incorporated into the chi-square distance measure to identify the writers as shown in equation 3 during the classification stage. In equation 3, N is the number of prototypes that are being clustered on an alphabet basis, where N=10 in this case.

The *AIC* takes into consideration the number of characters present in both the reference and test documents. This ensures that alphabets which seldom appear do not create distortions to the *tf-idf* distributions since an inadequate amount of characters might not be sufficient enough to model the writing style that is consistent with that alphabet. Hence, the *AIC* removes such bias by giving less significance to alphabets which are not frequent enough to completely model that alphabet.

3. Experimental Results

3.1. Database

Online handwritten documents have been collected with a digital pen and paper technology from 200 writers, where each writer has to copy two given text passages taken from the Reuters-21578 financial news corpus [13]. Hence one text passage is taken as a reference document and the other is taken as a test document. These reference and test documents that have been collected belong to a separate dataset from the IRONOFF database. The basis for using another dataset is that the IRONOFF database contains only isolated words and hence is not representative of actual online documents. Furthermore, this independence with the IRONOFF database allows a generic set of prototypes to be built from the IRONOFF database with respect to the actual reference documents. Figure 1 illustrates a typical online handwritten sample that has been collected.

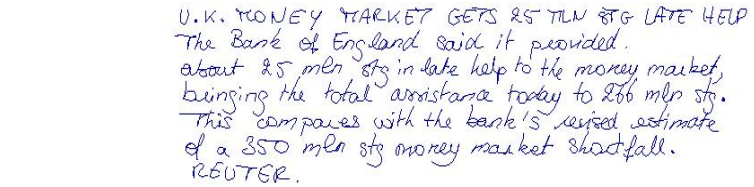

Figure 1. Typical handwritten text in database.

As can be seen from figure 1, each online document consists of only 3 to 10 lines of text and contains shorthand that typically exists in financial news. For example, mln which stands for million, is one of the commonly used financial words. The automatic recognizer tool might confuse the word 'mln' for 'men', thereby segmenting it wrongly to be 'e' due to this recognition error. A character recognition rate of 89% was achieved on this database using the MyScript [10] industrial text recognition engine. In the previous example that illustrated the incorrect segmentation of the word 'mln', a 'l' which is wrongly recognized as an 'e' will be erroneously clustered to the set of prototypes for alphabet 'e', thus corrupting the estimation of the frequency vectors with noise.

3.2. Impact of Alphabet Knowledge

Table 1. Impact of alphabet knowledge on accuracy.

Top-1 Writer Identification Rate		
Without Using Alphabet Knowledge	Using Alphabetic Knowledge	
	Without using AIC	With AIC
66.0%	73.5%	87.0%

Table 1 compares the top-1 writer identification rate between the approach without using any alphabet knowledge and that with alphabet knowledge. When alphabet knowledge was not used, an accuracy of 66.0% was achieved (68 misclassified writers out of 200). In contrast, when alphabet knowledge (without *AIC*) was used, we attain a higher top-1 identification of 73.5%. This translates into a misclassification error of only 53 out of 200 writers in the top-1 position. Furthermore, when we used *AIC*, our results further increase to 87%. Therefore, our results provide evidence that alphabet knowledge derived from the character prototypes contains valuable information on the writer, which allows the writer identification rate to be improved.

As explained in section 2.2, the *AIC* takes into account whether certain alphabets have been sufficiently modeled from the amount of characters present. Certain alphabets that rarely appear in the English language such as 'q' and 'z' [14] might impose a bias to the distribution built. The *AIC r*educes this bias by placing less emphasis on such infrequent alphabets that are unable to reliably model the prototypes. Therefore the introduction of the *AIC* can significantly improve the identification of writers.

3.3. Discriminative power of alphabets

Certain alphabets allow for more variations to be written compared with other alphabets, thereby allowing different writers to express their individuality of handwriting with a style that is more distinctive and differentiated. For example, the alphabet 'f' has more morphological variations and styles in its approach of writing compared to the alphabet 'c', where only a limited number of variations exist. This implies that most writers might inadvertently write the alphabet 'c' with a similar style. Therefore, we hypothesize that different alphabets will have different capabilities in identifying writers and we refer to this term as the discriminative power of alphabets. Experiments were conducted to verify this hypothesis by using only one alphabet at a time to identify the writers. For example, in order to investigate the discriminative power of alphabet 'a', only the alphabet 'a' was used from the reference and test documents in the document indexing stage and the retrieval stage. The top-1 accuracy in writer identification was then obtained by considering only writers that have the alphabet 'a' in both their reference and test documents. Writers that do not have any alphabet 'a' in either the reference or test document are omitted in the ranking results. This process is then repeated for the other alphabets. Four alphabets, namely, 'z', 'q', 'x' and 'j' were omitted for

the purpose of this experiment since these alphabets rarely appear in the documents and will skew the results if included. The outcome of this experiment is illustrated in table 2.

Table 2. Discriminative power of alphabets.

Alphabet	Top-1 Accuracy	% of total characters in documents
d	22.45%	4.45%
a	20.81%	8.17%
n	17.00%	7.41%
r	17.00%	7.21%
o	16.50%	7.89%
s	14.50%	7.64%
t	14.00%	8.67%
g	13.66%	1.57%
k	11.88%	0.73%
y	11.76%	1.47%
e	11.00%	12.17%
i	11.00%	7.55%
p	9.33%	2.31%
h	8.63%	3.39%
f	7.94%	2.21%
b	7.79%	1.37%
l	7.54%	4.20%
m	6.45%	2.44%
u	6.03%	3.16%
v	5.81%	0.95%
w	5.44%	1.16%
c	4.64%	3.25%

From table 2, the second column indicates the top-1 accuracy obtained when only that particular alphabet is used in performing writer identification. The results supported our hypothesis that certain alphabets like 'c' might have fewer variations of writing and hence will have a low discriminative power in writer identification. Likewise, alphabets like 'a' and 'd' are highly discriminative in writer identification. Results reported by our previous works [5] on French documents are similar to the results obtained here for English documents; with the notable exceptions that 's' and 't' are more discriminating in the French documents (top-2 and top-4 respectively in [5]) and that 'b' was the least discriminating in [5]. These slight differences are due to the fact that English and French documents are still inherently different. Nonetheless, this experiment clearly shows that different alphabets have different identification capabilities, which also supports findings from Cha et al. [15, 16]. Therefore more emphasis should be placed on such alphabets with high discriminative powers and less emphasis on those with low discriminative powers. The third column of table 2 shows the frequency of occurrence of such alphabets in both the test and reference documents. This distribution of alphabet frequency is similar to the results obtained by Foster [14] based on the Brown Corpus of US English words, where the most frequent alphabet is 'e', 't', 'a' and the least frequent alphabets are 'q' and 'z'.

In many pattern recognition scenarios, it is often imperative that one of the critical tasks consist of reducing redundancy in high dimensional feature spaces. The method proposed in [5] involves clustering in a feature space of dimensionality 210 for each alphabet, compounded by the fact that this has to be repeated for all 26 alphabets. This issue can be addressed by using only a subset of alphabets instead. We demonstrate the feasibility of using the discriminative power of the alphabet to determine a subset of alphabets. Figure 2 shows the drop in performance as alphabets are removed based on the order of their discriminative power. The experimental results show that the performance remains constant as the least discriminating alphabets are removed ('c', 'w', 'v', etc). Conversely, we suffer a drastic drop in the performance as the most discriminating alphabets are removed ('d', 'a', 'n', etc). Degradation in the performance is already observed even with the removal of one discriminating alphabet. Our results indicate that a choice of alphabets based merely on their discriminative power can be utilized to select a subset of alphabets without adverse impact to the performance of the writer identification system. This will help to reduce the dimensionality and decrease the computational complexity of the system. As evidenced from our results, this method of using the discriminative power to select a subset, albeit sub-optimal, is a much simpler and effective approach compared to other more complex and time-consuming subset selection algorithms.

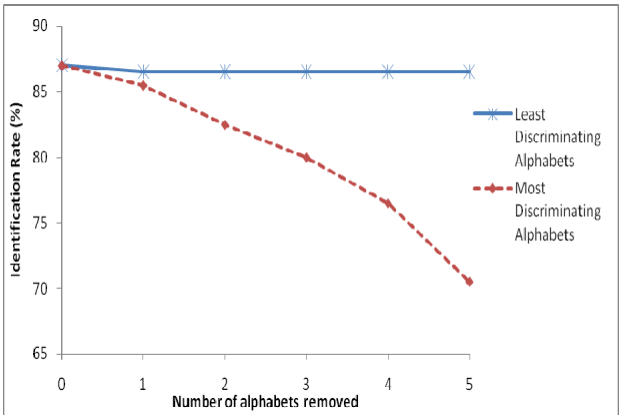

Figure 2. Identification rate as alphabets are removed based on their discriminative power.

4. Discussions

Alphabet knowledge contains valuable information pertaining to the identity of the writer. This is demonstrated from the experimental results where we are able to improve the performance from 66% to 87% by considering alphabet knowledge. Such alphabet knowledge can only be found if the prototype templates are built at the character level. The identification rate reported here are much lower compared to [5], although the reader should bear in mind that the Reuters database is much more challenging to recognize due to the shorter length of text in the documents and a higher segmentation error. Nonetheless, the emphasis of this paper is to investigate the impact of alphabet knowledge on writer identification. We have demonstrated from our experimental results that building prototype templates at the character level retains significant information relevant to the writer, where the alphabet knowledge can be inferred from the character prototypes. Therefore, alphabet knowledge helps in the identification of writers by allowing certain alphabets that are more relevant to certain writers to be closely examined. For this reason, the results presented in this paper explain why works that made use of character prototype approaches for writer identification are able to attain promising results.

We also establish the notion of the discriminative power of alphabets and the feasibility in utilizing such information towards reducing the dimensionality and complexity of the writer identification. Even though the concept of using the discriminative power of alphabets has been shown to be feasible, some open issues remain. One underlying assumption here is that the alphabets are independent of one another. This might not necessarily be the case if co-dependencies between alphabets exist. Hence, future work will be to take such co-dependencies into account when selecting a subset based on the discriminative power. Another open issue will be how to adapt this discriminative power of alphabets to specific writers. This way, we will be able to determine alphabets which can be ignored for those writers based on the discriminative power of alphabets for that specific writer. These issues will be addressed in our future work.

Acknowledgements

This research is jointly supported by Nanyang Technological University RSS Grant in Singapore, the French Merlion Scholarship and the ANR Grant CIEL 06-TLOG-009.

Reference

[1] R. A. Huber and A. M. Headrick, *Handwriting Identification - Facts and Fundamentals*: CRC Press, 1999.

[2] R. N. Morris, *Forensic Handwriting Identification - Fundamentals, Concepts and Principals*: Academic Press, 2000.

[3] M. Bulacu and L. Schomaker, "Text-independent writer identification and verification using textural and allographic features," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 29, pp. 701-717, Apr 2007.

[4] R. Niels, F. Gootjen, and L. Vuurpijl, "Writer Identification through Information Retrieval: The Allograph Weight Vector," in *International Conference on Frontiers in Handwriting Recognition*, 2008, pp. 481-486.

[5] G. X. Tan, C. Viard-Gaudin, and A. Kot, "Automatic Writer Identification Framework for Online Handwritten Documents Using Character Prototypes," *Pattern Recogn.,* 2009.

[6] L. Schomaker and M. Bulacu, "Automatic writer identification using connected-component contours and edge-based features of uppercase western script," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 26, pp. 787-798, 2004.

[7] A. Bensefia, T. Paquet, and L. Heutte, "Handwritten Document Analysis for Automatic Writer Recognition," *Electronic Letters on Computer Vision and Image Analysis,* vol. 5, pp. 72-86, 2005.

[8] L. Schomaker, "Advances in Writer Identification and Verification," in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, 2007, pp. 1268-1273.

[9] C. Viard-Gaudin, P. M. Lallican, S. Knerr, and P. Binter, "The ireste on/off (ironoff) dual handwriting database," *IEEE Int. Conf. Document Analysis and Recognition,* pp. 455-458, 1999.

[10] "Vision Objects Industrial Text Recogniser SDK MyScript Builder Help," in *SDK documentation*: http://www.visionobjects.com/about-us/download-center/_263/myscript-products-datasheets.html, 2009.

[11] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM,* vol. 18, pp. 613-620, 1975.

[12] J. Han and M. Kamber, *Data Mining - Concepts and Techniques*: Elsevier, 2006.

[13] S. P. Saldarriaga, C. Viard-Gaudin, and E. Morin, "On-line Handwritten Text Categorization," in *SPIE-IS&T Electronic Imaging: Document Recognition and Retrieval XVI*, 2009.

[14] C. C. Foster, *Cryptanalysis for microcomputers*: Hayden Book Co, 1982.

[15] S. Cha, S. Yoon, and C. C. Tappert, "Handwriting Copybook Style Identification for Questioned Document Examination," *Journal of Forensic Document Examiners,* pp. 1-14, 2007.

[16] S. N. Srihari, S. H. Cha, H. Arora, and S. Lee, "Individuality of handwriting," *Journal of Forensic Sciences,* vol. 47, pp. 856-872, Jul 2002.